Tutorial for Gene Expression Analysis: Identification Of Candidate Anti-Cancer Molecular Mechanisms Of Compound Kushen Injection Using Functional Genomics

BIOTECH-7005-BIOINF-3000

Zhipeng Qu

School of Biologial Sciences
The University of Adelaide

11/10/2022

Outline

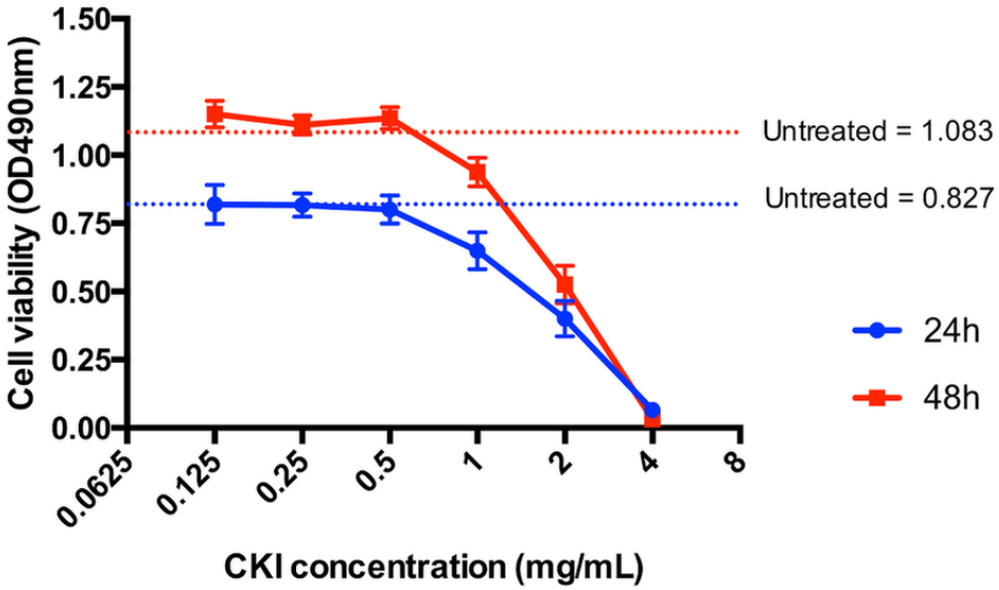
- Research background
- Differential gene expression analysis
  - High-throughput sequencing
  - QC
  - Genome mapping
  - Count reads
  - DE analysis
- Experimental validation
- Conclusions

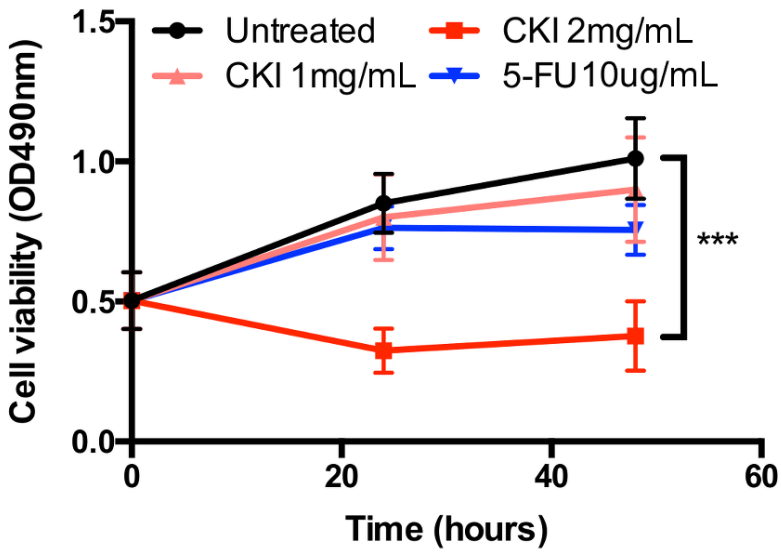
Research background

Research background

- CKI (Compound Kushen Injection)
  - Kushen (*Radix Sophorae Flavescentis*) and Baituling (*Rhizoma smilacis Glabrae*)
  - Alkaloids (matrine, oxymatrine and sophoridine ...)
- Anti-tumor effect (based on matrine/oxymatrine studies)
- Cancer pain relief (Z Zhao et al., Cancer letters, 2014)

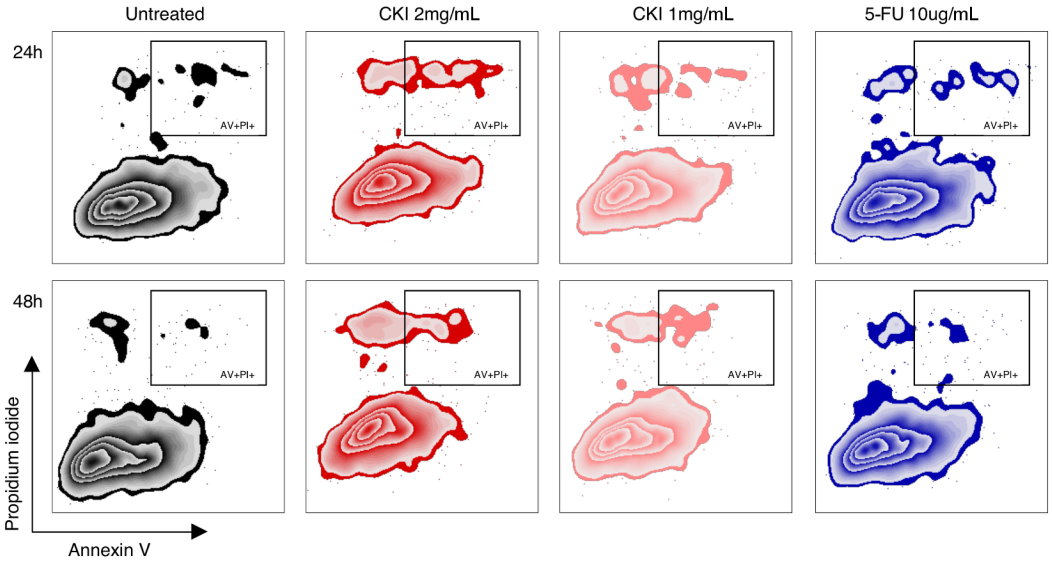
CKI inhibits MCF-7 cell proliferation

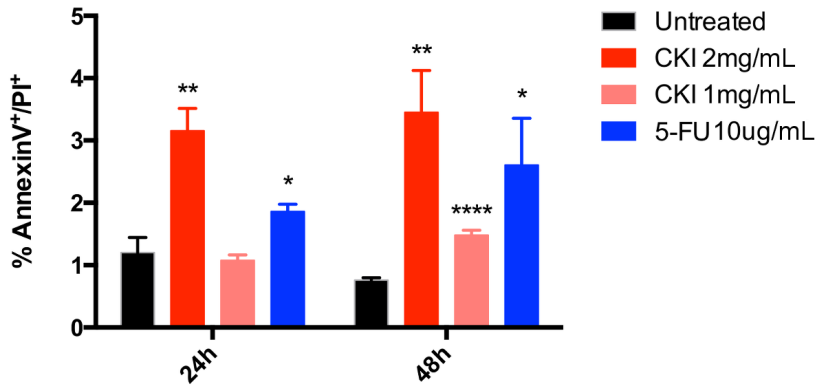




Cell viability assay showing CKI can inhibit MCF-7 cell proliferation in a dose-dependent fashion.

CKI induces MCF-7 cell apoptosis





Annexin V/PI assay showing CKI can induce apoptosis in MCF-7 cells.

Research questions?

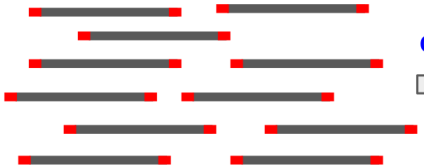
- Whether there are genes affected by CKI in cancer cells (Differential expression between CKI-treated cancer cells and control cancer cells)?
- If there are, what are those affected genes (DE genes)?
- What are the molecular functions of those DE genes and what are the molecular pathways that these DE genes involved in?

Differential gene expression analysis

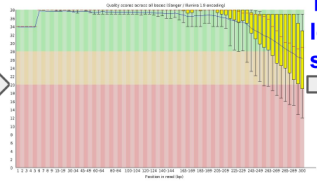
Overview of DE analysis

Base calling

Raw reads

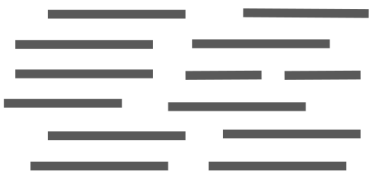


QC

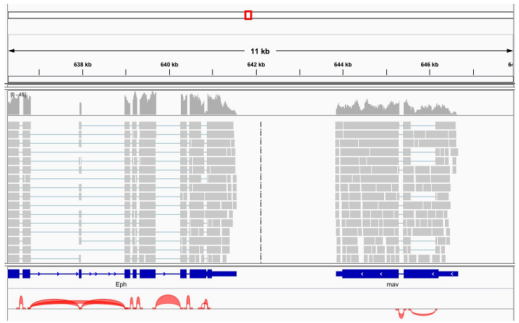


Trim adaptors and low quality sequences

Clean reads

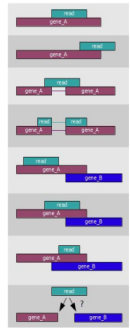


Genome mapping

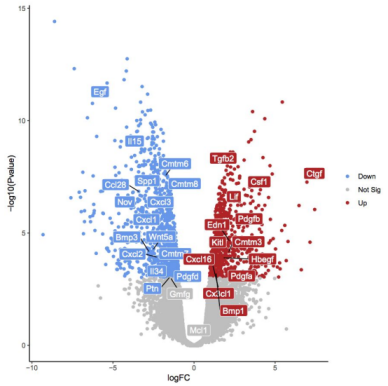


Feature count

	Sample1	Sample2	...
Gene 1	56	209	...
Gene 2	89	10	...
Gene 3	0	503	...
Gene 4	20	200	...
Gene 5	2	20	...
...	...	...	...

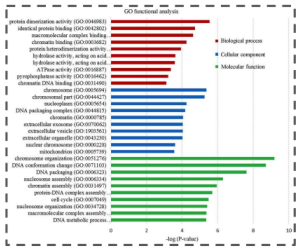


DE gene identification

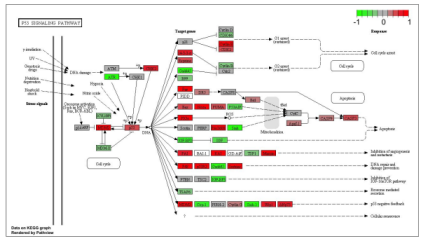


Functional analysis

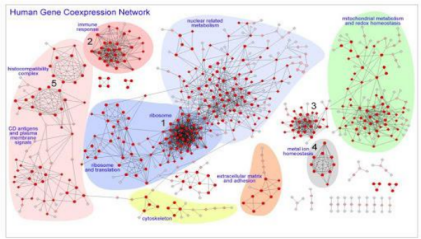
GO enrichment analysis



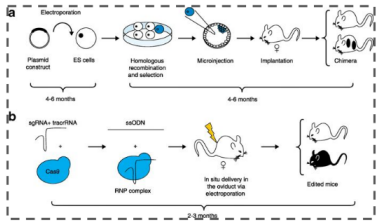
KEGG functional annotation



Co-expression networks



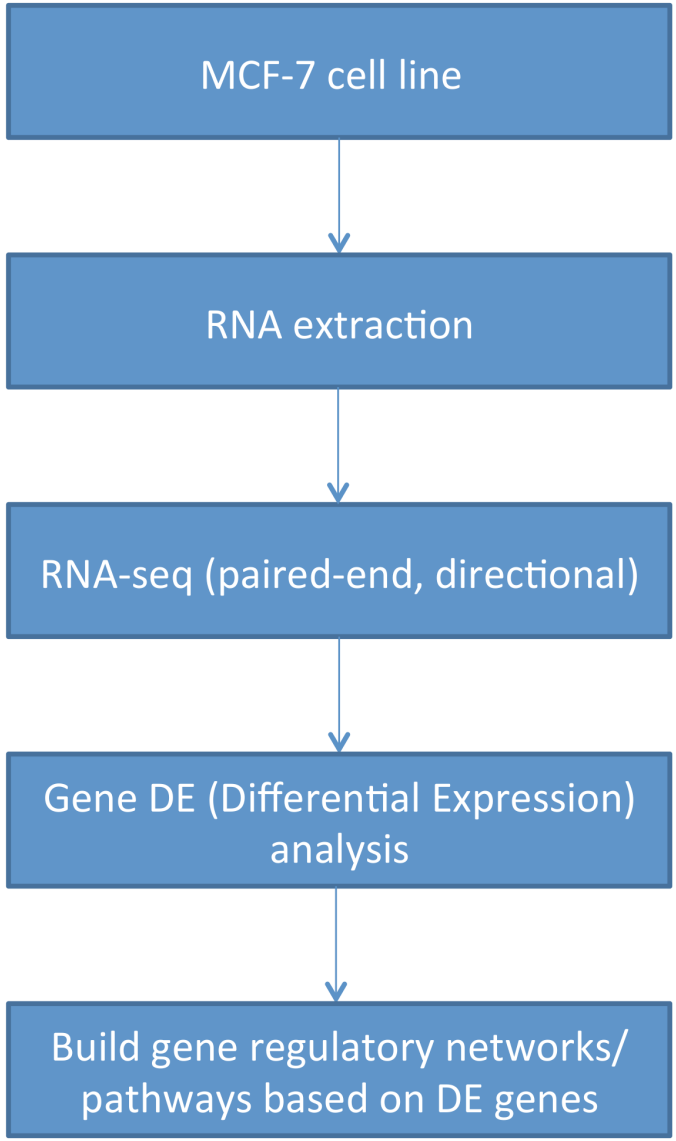
Functional validation



Transcriptome project summary

- Time points: 0 hrs, 24 hrs, 48 hrs
- 3 groups (3 replicates)
  - control
  - CKI
  - positive control 5-FU (Fluorouracil)*
- 2 doses for CKI (final concentration in culture medium)**
  - 1 mg/mL
  - 2 mg/mL

* 5-FU is a thymidylate synthase (TS) inhibitor; induces cell death via lack of thymine
** The dose of CKI is measured based on the concentration of total alkaloids



Sequencing statistics

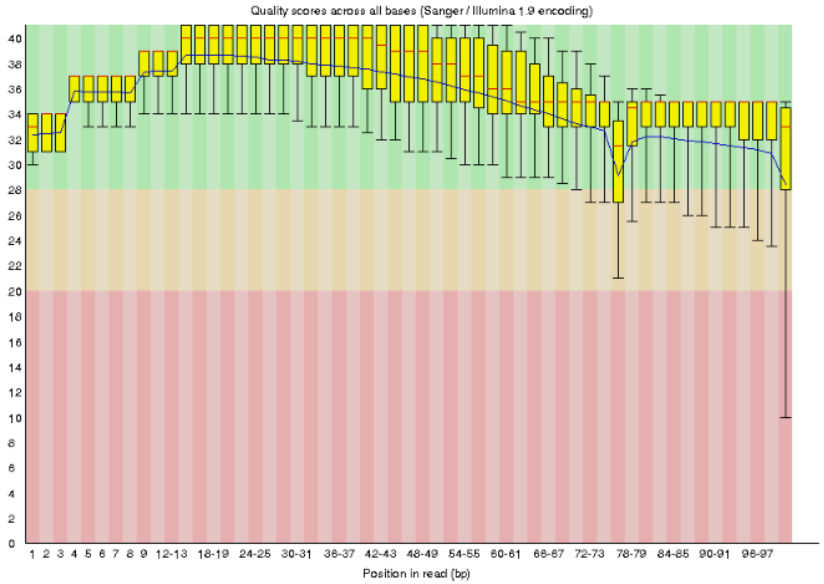
- Sequencing platform: Illumina
- Sequencing type: PE100 (paired-end 100bp)
- Sequencing coverage: 20-30 million read pairs per sample
- Number of sequenced samples: 21 samples

Before trimming

Basic Statistics

Measure	Value
Filename	0h_Ctrl_R2_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	27113173
Sequences flagged as poor quality	0
Sequence length	101
%GC	55

Per base sequence quality

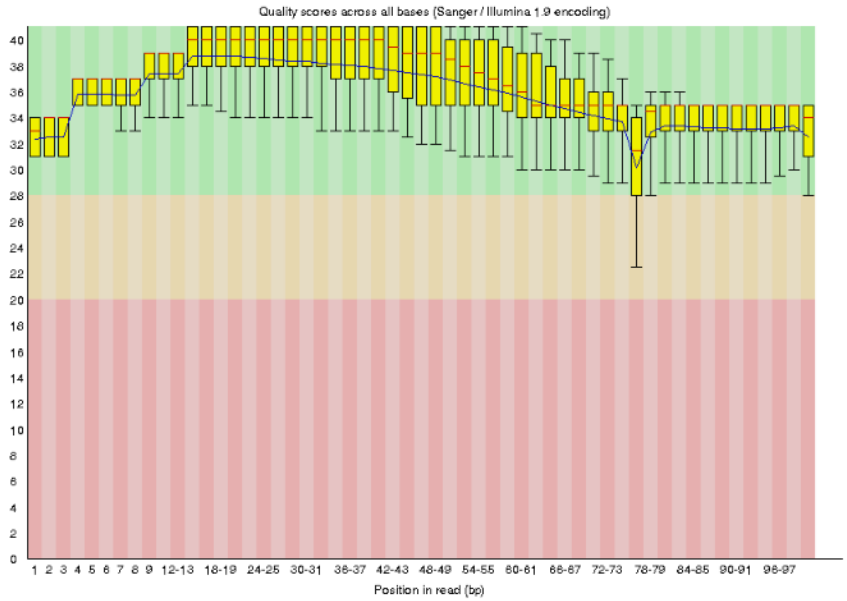


After trimming

Basic Statistics

Measure	Value
Filename	0h_Ctrl_R2_R1_val_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	26530224
Sequences flagged as poor quality	0
Sequence length	20-101
%GC	55

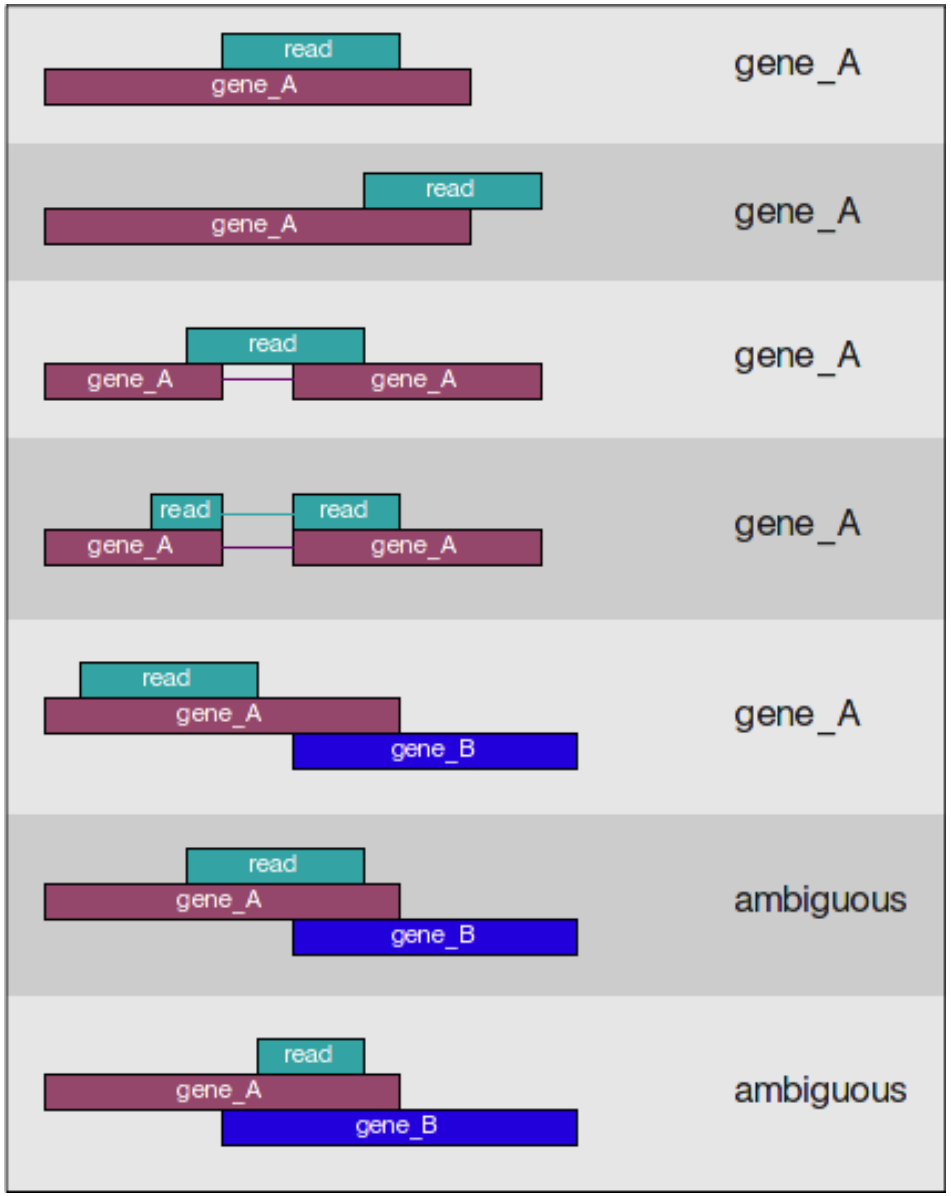
Per base sequence quality



Genome mapping

CCND1: encoding Cyclin D1 protein, important for cell cycle G1/S transition

Read counting



Gene ID	X0h_Ctrl_R2	X0h_Ctrl_R4	X0h_Ctrl_R5
A1BG	81	90	119
A1BG-AS1	483	249	326
A1CF	0	0	0
A2M	45	79	0
A2M-AS1	30	14	7
A2ML1	0	64	0
A2MP1	1	15	0
A3GALT2	0	0	0
A4GALT	2568	2916	2317
A4GNT	0	0	0
AA06	0	0	0
AAAS	3121	3718	2620
AACS	2602	2807	2130
AACSP1	0	0	0
AADAC	0	0	0
AADACL2	0	0	0
AADACL2-AS1	0	0	0
AADACL3	0	0	0
AADACL4	0	0	0
AADACP1	0	0	0
AADAT	70.75	42	19
AAED1	81	146	17
AAGAB	924	1893	370
AAK1	429	365	152
AAMDC	140	258	130
AAMP	5918	6804	5981
AANAT	0	0	0
AAR2	1346	1300	1177
AARD	234	304	135
AARS	8673	9592	5784
AARS2	709	854	392
AARSD1	435	465	378
AASDH	13	0	0
AASDHPPT	42	120	35
AASS	0	0	0

DE analysis using edgeR

BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 1 2010, pages 139–140
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Received on March 29, 2009; revised on October 19, 2009; accepted on October 23, 2009

Advance Access publication November 11, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: It is expected that emerging digital gene expression (DGE) technologies will overtake microarray technologies in the near future for many functional genomics applications. One of the fundamental data analysis tasks, especially for gene expression studies, involves determining whether there is evidence that *counts* for a transcript or exon are significantly different across experimental conditions. `edgeR` is a Bioconductor software package for examining differential expression of replicated count data. An overdispersed Poisson model is used to account for both biological and technical variability. Empirical Bayes methods are used to moderate the degree of overdispersion across transcripts, improving the reliability of inference. The methodology can be used even with the most minimal levels of replication, provided at least one phenotype or experimental condition is replicated. The software may have other applications beyond sequencing data, such as proteome peptide count data.

Availability: The package is freely available under the LGPL licence from the Bioconductor web site (http://bioconductor.org).

Contact: mrobinson@wehi.edu.au

(SAGE), the methods and software should be equally applicable to emerging technologies such as RNA-seq (Li *et al.*, 2008; Marioni *et al.*, 2008) giving rise to digital expression data. `edgeR` may also be useful in other experiments that generate counts, such as ChIP-seq, in proteomics experiments where spectral counts are used to summarize the peptide abundance (Wong *et al.*, 2008), or in barcoding experiments where several species are counted (Andersson *et al.*, 2008). The software is designed for finding changes between two or more groups when at least one of the groups has replicated measurements.

2 MODEL

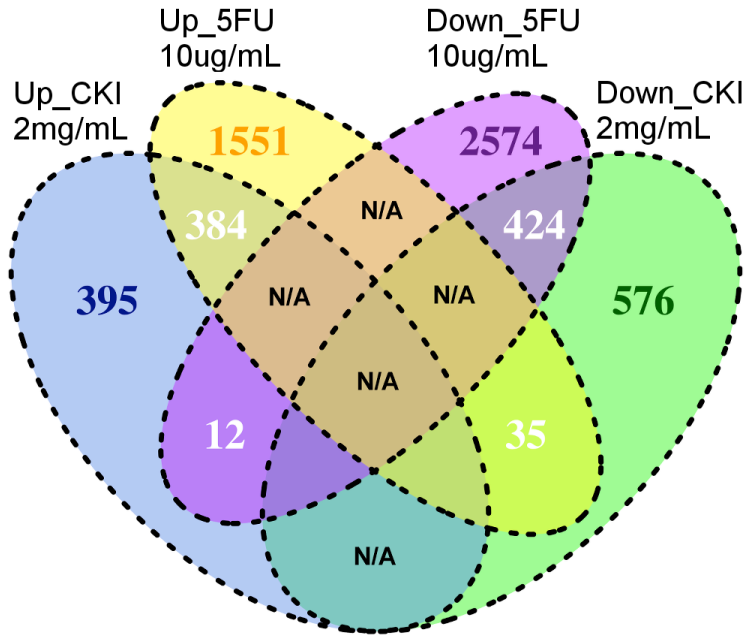
Bioinformatics researchers have learned many things from the analysis of microarray data. For instance, power to detect differential expression can be improved and false discoveries reduced by sharing information across all probes. One such procedure is `limma` (Smyth, 2004), where an empirical Bayes model is used to moderate the probe-wise variances. The moderated variances replace the probe-wise variances in the *t*- and *F*-statistic calculations. In a closely

Output of DE analysis using edgeR

Gene ID	logFC	logCPM	Pvalue (Adjusted)
CYP1A1	6.795614577	8.620629982	2.57E-72
HMOX1	4.893276588	7.182980975	2.89E-37
MAOB	-4.157864167	5.029132457	1.83E-31
IL20	-7.10491203	4.018033996	4.81E-29
CCND1	-2.54801286	8.946777993	1.00E-21
AKR1C2	2.738849661	6.232611749	2.19E-20
H19	-3.226924218	6.779824345	7.49E-20
AKR1C3	4.413551172	4.377842126	3.50E-19
THBS1	-3.282170203	8.153376999	2.65E-18
TMPRSS2	3.503863685	4.857941994	1.43E-17
ASCL1	-4.418408568	4.520828682	7.18E-17
MCM3	-2.2400994	6.236342496	1.45E-16
OAS3	-2.64444225	6.562771277	2.30E-16
GFRA1	-1.887691947	7.815781191	1.04E-15
PLXNA4	-4.162450018	4.770524947	2.15E-15
UHRF1	-3.150385396	5.775519462	3.25E-15
DUSP13	4.233207685	4.078589934	1.40E-14

Summary of differential gene expression analysis (24hr)

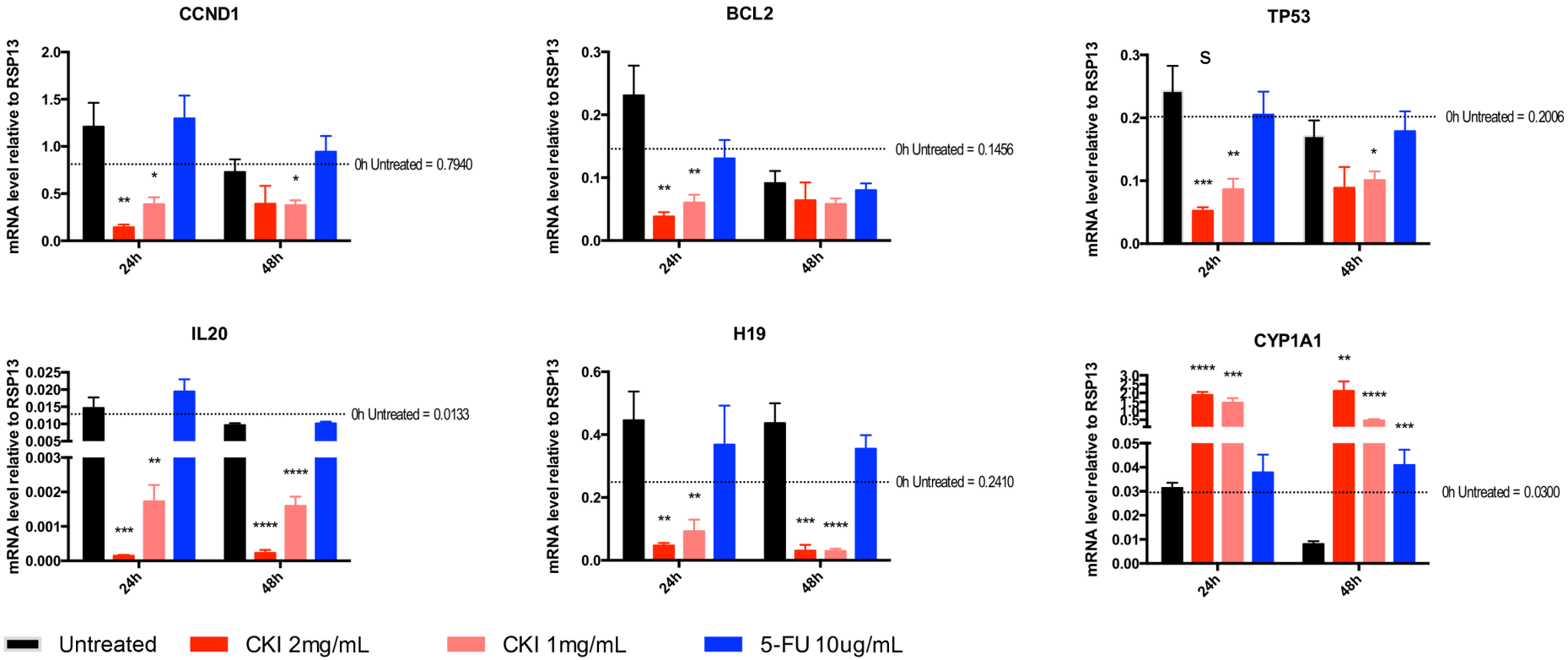
24h	Untreated	CKI (1mg/mL)	CKI (2mg/mL)	5-FU (10ug/mL)
Untreated		–62	–1035	–3010
CKI (1mg/mL)	85		–56	–2231
CKI (2mg/mL)	791	5		–1877
5-FU (10ug/mL)	1970	1264	1348	



Left panel, Numbers of significantly differentially expressed genes between different groups (row names compared to column names) (edgeR, FDR < 0.05). "yellow colour" represents up-regulation and "blue colour" represents down-regulation. **Right panel**, Comparison of significantly differentially expressed genes between MCF-7 cells treated with CKI or 5-FU compared to untreated cells.

What to do next after you get DE genes?

Validation of important differentially expressed genes with qPCR

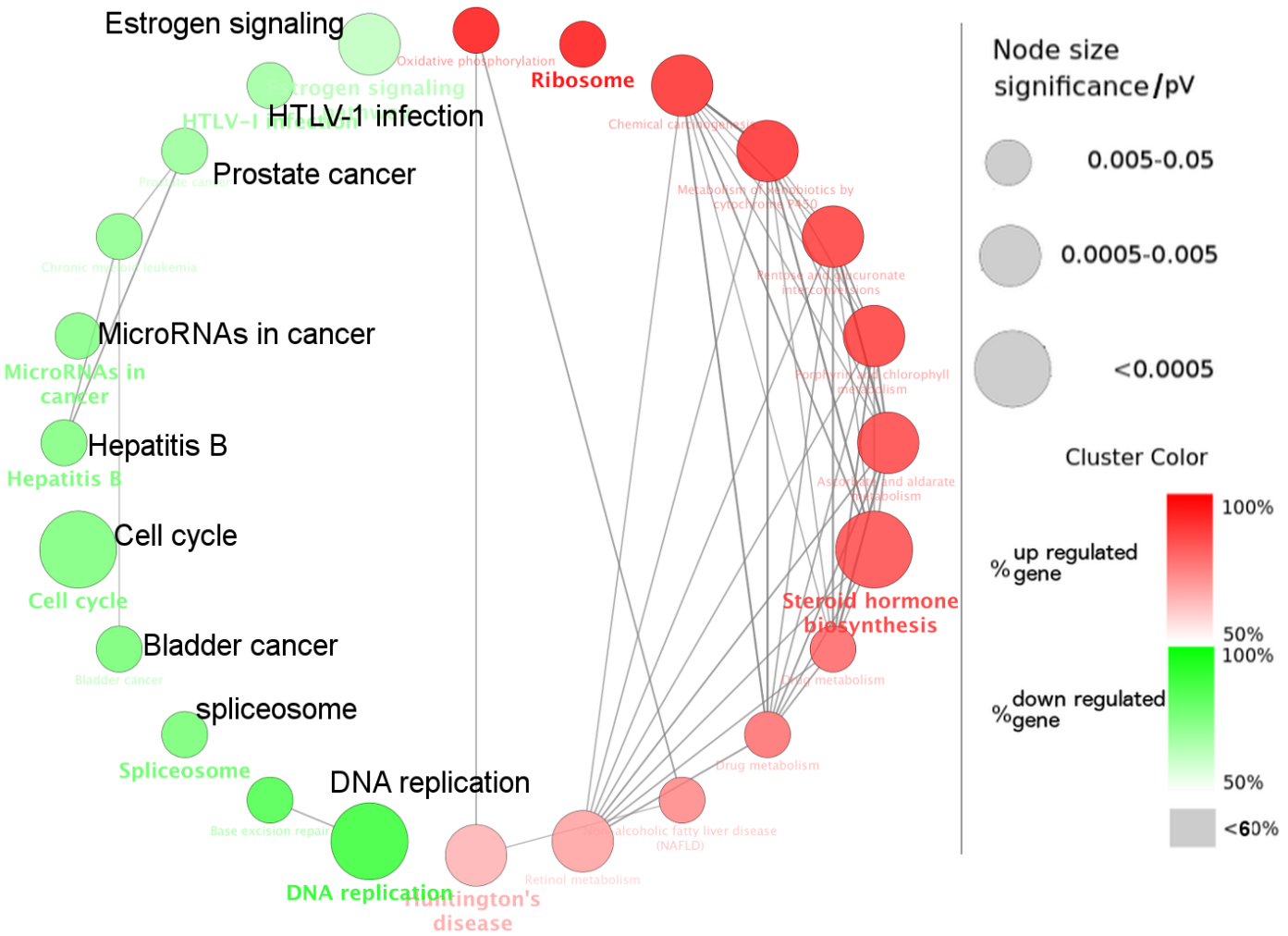


	Gene	Full name	Function
1	*CCND1*	Cyclin D1	protein kinase activity, cell cycle
2	*BCL2*	B-Cell CLL/Lymphoma 2	Oncogene, anti-apoptotic
3	*TP53*	Tumor Protein P53	Tumour suppressor, cell division
4	*IL20*	Interleukin 20	cytokine activity and interleukin-20 receptor binding
5	*H19*	Imprinted Maternally Expressed Transcript (Non-Protein Coding)	Beckwith-Wiedemann Syndrome and Wilms tumorigenesis, over-expressed in breast cancer
6	*CYP1A1*	Cytochrome P450, Family 1, Subfamily A	enzyme binding and iron ion binding

GO over-representation analysis: CKI

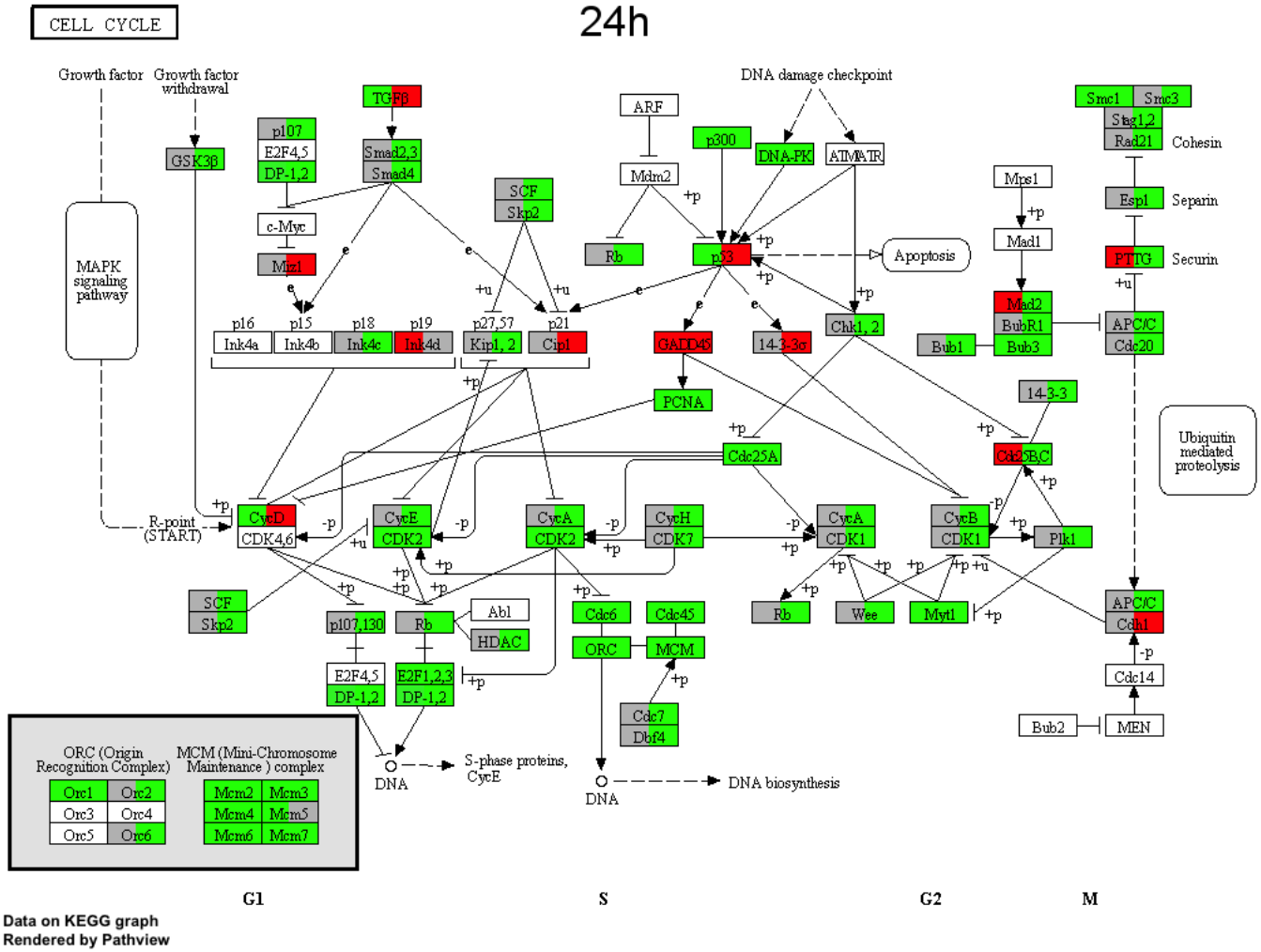
Over-represented GO (Gene Ontology) terms (biological process) for all significantly differentially expressed genes between cells treated with 2mg/mL CKI for 24 hours compared to untreated cells.

KEGG pathway over-representation analysis: CKI



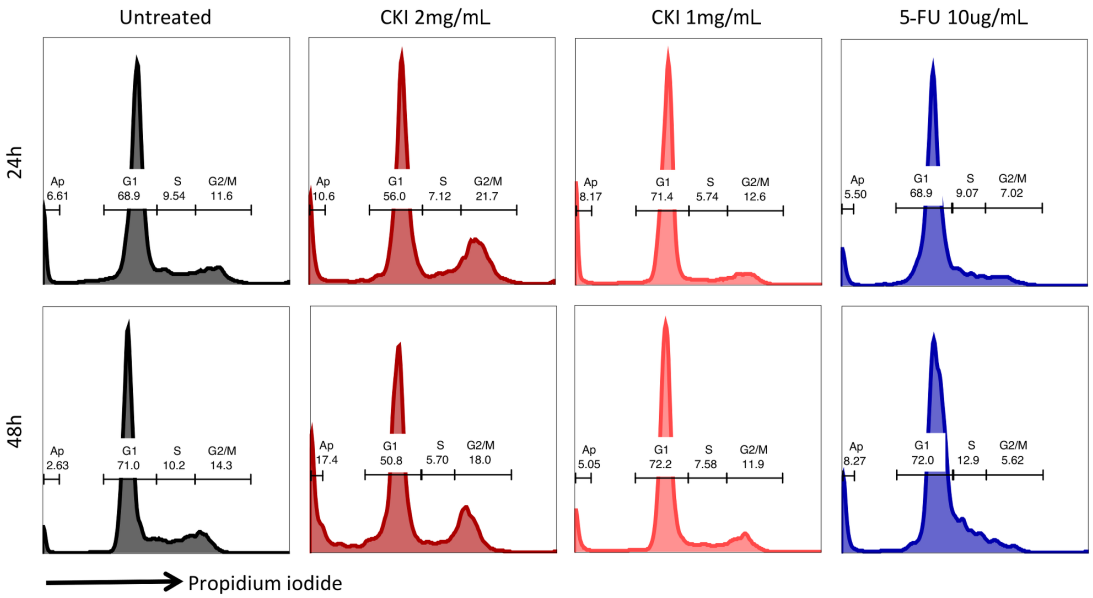
Over-represented KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways for all significantly differentially expressed genes between cells treated with 2mg/mL CKI for 24 hours compared to untreated cells.

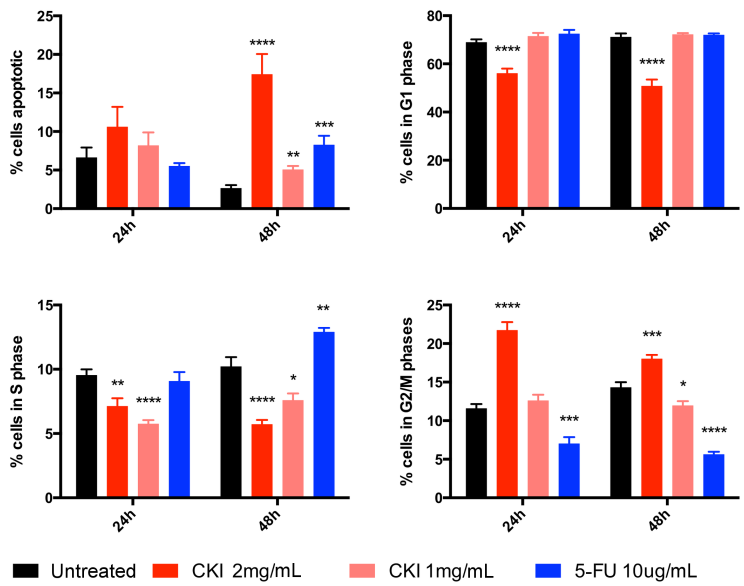
Cell cycle as a primary target molecular pathway of CKI in MCF-7 cells



Gene expression changes in the cell cycle pathway in MCF-7 cells treated with 2mg/mL CKI (**left half**) or 5-FU (**right half**) for 24 hours compared to untreated cells. Red colour: up-regulation; Green colour: down-regulation.

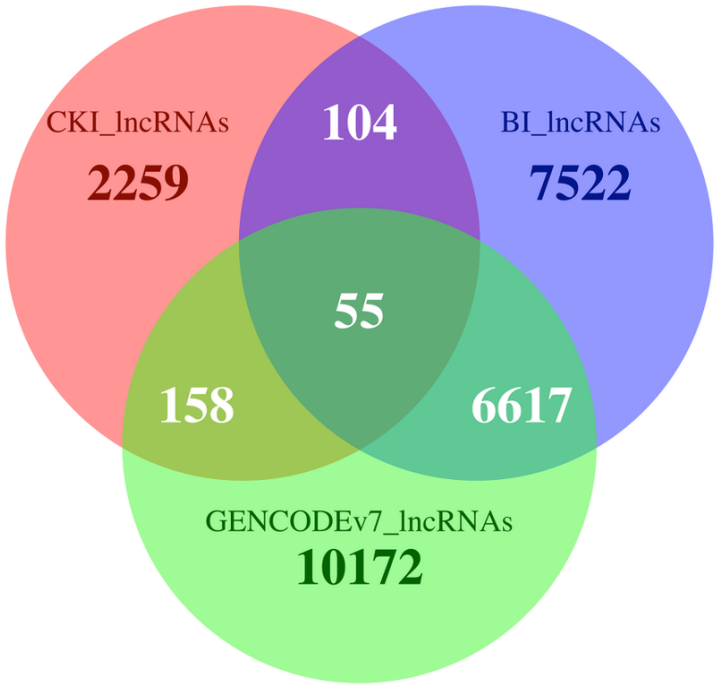
Cell cycle as a primary target molecular pathway of CKI in MCF-7 cells

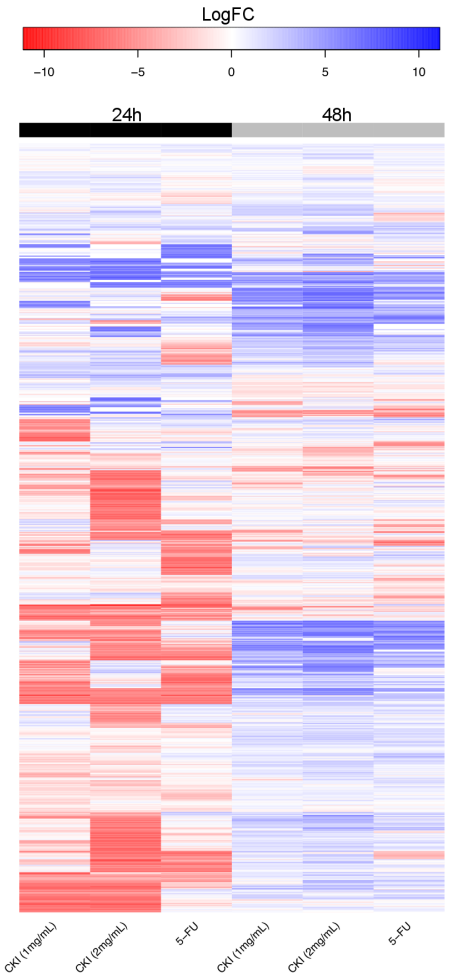




Cell cycle assay indicating a cell cycle arrest at G2/M phase induced by CKI in MCF-7 cells.

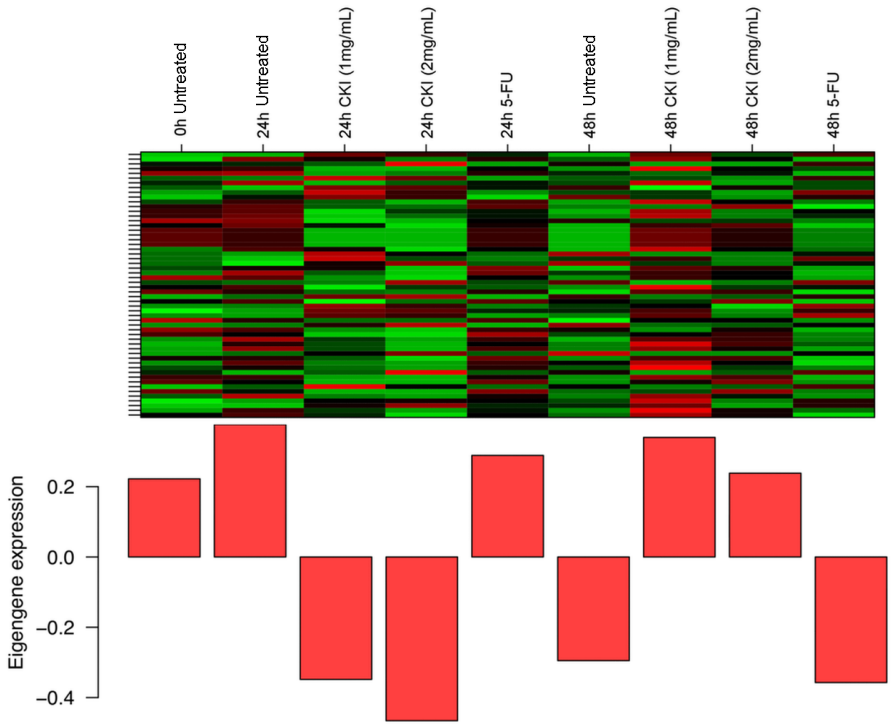
Expression of many lncRNAs shows response to CKI treatment in MCF-7 cells

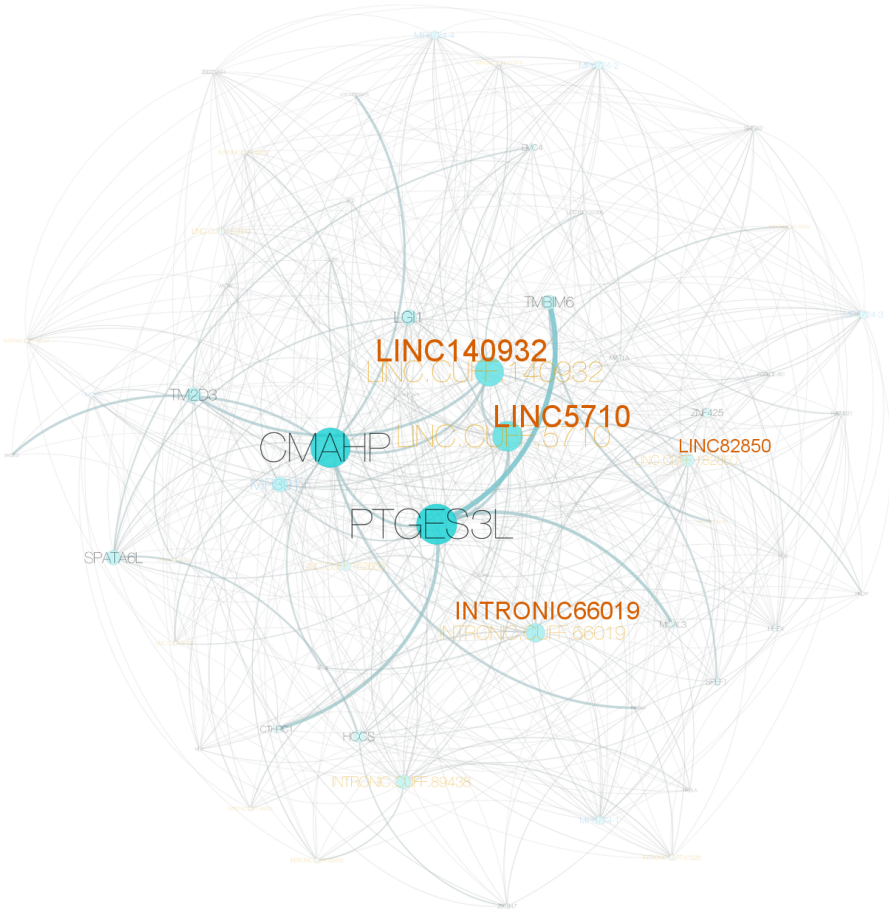




More than 2000 lncRNAs are *de novo* identified from transcriptome (**left panel**), and many of them might be expressed as response to CKI treatment (**right panel**).

Genes and lncRNAs are co-expressed as a response to CKI treatment





A co-expression module/sub-network shows response to CKI treatment (**left panel**), and some lncRNAs (nodes in orange colour) were identified as hubs of these co-expression module (**right panel**).

Functional annotation of protein-coding genes in CKI-specific co-expression module

	Term	Fold enrichment	P-value
1	GO:0008283 cell proliferation	3.354	0.009
2	GO:0010604 positive regulation of macromolecule metabolic process	2.346	0.016
3	GO:0032989 cellular component morphogenesis	3.223	0.02
4	GO:0048514 blood vessel morphogenesis	4.332	0.027
5	GO:0035295 tube development	4.155	0.031
6	GO:0007242 intracellular signaling cascade	1.892	0.036
7	GO:0019932 second-messenger-mediated signaling	3.89	0.038
8	GO:0000904 cell morphogenesis involved in differentiation	3.746	0.043
9	GO:0000902 cell morphogenesis	3.081	0.043
10	GO:0001568 blood vessel development	3.731	0.043
11	GO:0001944 vasculature development	3.642	0.047
12	hsa05200:Pathways in cancer	3.391	0.013
13	hsa04020:Calcium signaling pathway	4.514	0.021

Over-represented GO or KEGG terms (DAVID, p-value < 0.05)

Tutorial summary

- Investigate your research background and ask your research question(s)
- Design your DE analysis properly
- Get DE genes
- Interpret your DE analysis results by additional analyses
- Functional validation using experiments